Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale

Year: 2023 | Citations: 412 | Authors: Matt Le, Apoorv Vyas, Bowen Shi, B. Karrer, Leda Sari

Abstract

Large-scale generative models such as GPT and DALL-E have revolutionized the research community. These models not only generate high fidelity outputs, but are also generalists which can solve tasks not explicitly taught. In contrast, speech generative models are still primitive in terms of scale and task generalization. In this paper, we present Voicebox, the most versatile text-guided generative model for speech at scale. Voicebox is a non-autoregressive flow-matching model trained to infill speech, given audio context and text, trained on over 50K hours of speech that are not filtered or enhanced. Similar to GPT, Voicebox can perform many different tasks through in-context learning, but is more flexible as it can also condition on future context. Voicebox can be used for mono or cross-lingual zero-shot text-to-speech synthesis, noise removal, content editing, style conversion, and diverse sample generation. In particular, Voicebox outperforms the state-of-the-art zero-shot TTS model VALL-E on both intelligibility (5.9% vs 1.9% word error rates) and audio similarity (0.580 vs 0.681) while being up to 20 times faster. Audio samples can be found in \url{https://voicebox.metademolab.com}.